Experiment Report 4: Hardware Comparison on Tiny Shakespeare (Character-Level nanoGPT)

Date: Tuesday, 02/17/2026
Participants: John Lee, Maximo Sanchez, Youssif Goda

Experimental Overview:

This study investigates whether differences in hardware specifications, specifically **CPU** versus **GPU**, and across GPU tiers to investigate if they produce observable differences in training behavior when scaling a character-level language model across three complexity levels: **Low**, **Medium**, and **Large**. To this end, a small-scale GPT model was trained on the Tiny Shakespeare dataset using the nanoGPT framework, with training monitored across a suite of performance and energy-efficiency metrics beyond conventional loss curves.

Beyond standard loss-based evaluation, this experiment tracks several hardware-sensitive metrics: tokens per second (throughput), power consumption in watts, and energy expenditure measured in joules, both cumulatively and per iteration. Of particular analytical interest are two composite graphs introduced to assess energy efficiency:

- **Total Tokens per Total Joules**: a cumulative efficiency graph plotting work done (tokens processed) against fuel consumed (joules expended).
- **Token Throughput vs. Joules per Iteration**: an efficiency-versus-speed graph relating instantaneous throughput (tokens per second) to energy cost per iteration (joules).

Experimental Setup:

All participants trained an identical model configuration using the nanoGPT framework within the shared repository: https://github.com/MaxiSanc37/GPU-vs-CPU-Comparative-Research. To ensure reproducibility and enable cross-hardware comparison, all runs were logged to a shared Weights & Biases project: *GPUvsCPUResearch*, organized by complexity group:

- **Low Complexity** (3 heads, 3 layers, 126-dim embedding): W&B group
- **Mid Complexity** (6 heads, 6 layers, 384-dim embedding): W&B group
- **Large Complexity** (12 heads, 12 layers, 768-dim embedding): W&B group

The experiment was conducted across **five** hardware configurations spanning three machines, ordered conceptually from highest to lowest theoretical compute capability:

1. NVIDIA RTX 4060 GPU
2. NVIDIA RTX 3070 GPU
3. Apple M2 Silicon (unified memory architecture)
4. AMD Ryzen 7 8845HS (CPU)
5. AMD Ryzen 7 5800H (CPU)

All models were trained for an identical number of iterations using fixed hyperparameters, random seeds, dataset splits, and a uniform **batch size of 12**, ensuring that any observed differences in behavior are attributable to hardware rather than configuration variance.

Results and Observations:

Across all hardware configurations and complexity levels, **training loss** and **validation loss** curves were nearly identical. This result is expected: because all runs shared the same model architecture, hyperparameters, dataset splits, and random seeds, the underlying optimization problem was unchanged, and learning dynamics remained consistent across devices. Loss convergence therefore serves as a control confirming experimental integrity rather than a differentiating metric.

To capture hardware-level differences, GPU energy metrics were collected using the `pynvml` library, which provides programmatic access to NVIDIA GPU telemetry. This enabled the logging of per-iteration and cumulative energy consumption in joules for the RTX 4060 and RTX 3070, quantifying the energy cost of token generation across training. Power consumption in watts was also recorded, though this metric had already surfaced through Weights & Biases, likely due to an existing integration within `train.py`.

The following metrics and visualizations were introduced to characterize hardware performance:

Runtime captures total wall-clock training duration as a function of iteration, providing a direct measure of hardware speed.

- X-axis: Iteration number
- Y-axis: Runtime (seconds)

Tokens Per Second measures instantaneous throughput, reflecting each device's capacity to process training data and serving as a hardware-normalized proxy for learning rate.

- X-axis: Iteration number
- Y-axis: Tokens per second

Total Tokens per Total Joules ("Efficiency over Time") is a cumulative graph that relates total work performed to total energy consumed, enabling comparison of energy efficiency trajectories across hardware over the full training run.

- X-axis: Cumulative tokens processed
- Y-axis: Cumulative joules consumed

Token Throughput vs. Joules per Iteration ("Efficiency vs. Speed") plots instantaneous energy cost against instantaneous throughput, revealing whether higher-performing hardware achieves its speed advantage efficiently or at disproportionate energy cost.

- X-axis: Tokens per second
- Y-axis: Joules per iteration

Note regarding energy consumption metrics:

(joules per iteration and cumulative joules) are exclusively available for the NVIDIA RTX 4060 and RTX 3070 configurations. This is a direct consequence of the data collection method: the `pynvml` library (NVIDIA Management Library Python bindings) interfaces with NVIDIA's proprietary GPU driver API, which exposes hardware-level power and energy telemetry in real time.

This interface is only available on NVIDIA discrete GPUs and is not supported on CPU-based systems or Apple Silicon. As a result, the AMD Ryzen 7 8845HS, AMD Ryzen 7 5800H, and Apple M2 Silicon configurations do not report joules data and are therefore absent from the energy-efficiency graphs (Total Tokens per Total Joules and Token Throughput vs. Joules per Iteration).

These hardware configurations are instead evaluated exclusively through the hardware-agnostic metrics such as runtime in seconds and tokens per second which are computed from system timestamps and iteration counts rather than hardware sensor readings, and are therefore universally available regardless of device type or vendor.

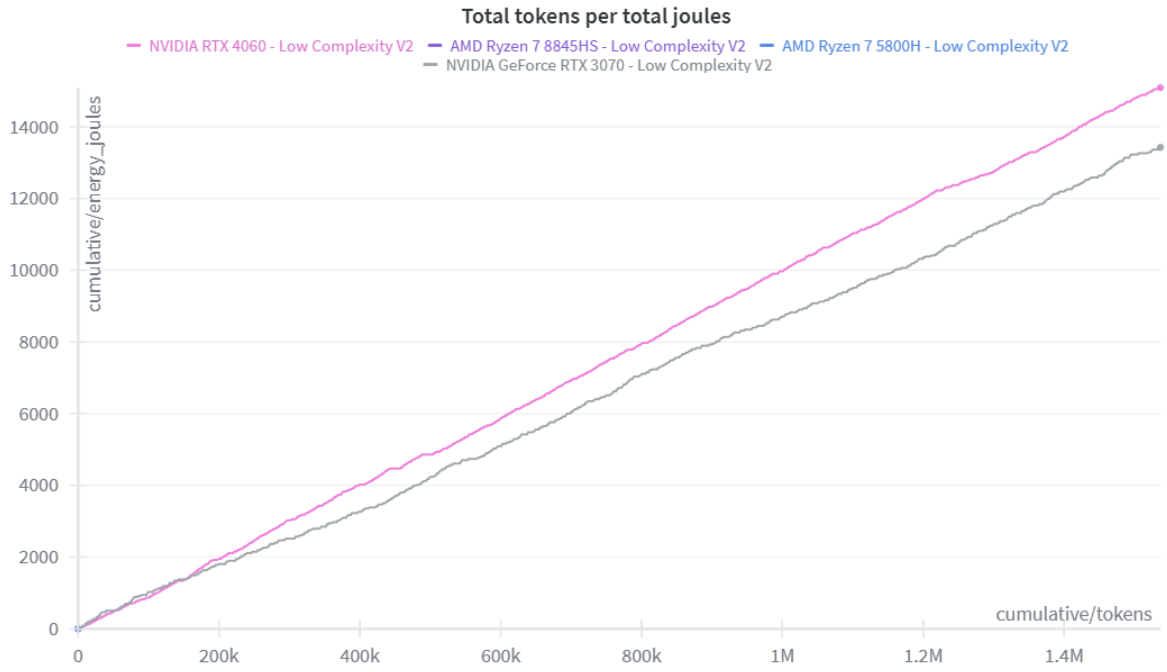
This asymmetry in metric availability is an inherent limitation of the current instrumentation approach, not an experimental oversight. Future work could address this gap through platform-agnostic power monitoring tools such as Intel RAPL (for x86 CPUs) or vendor-specific Apple Silicon profiling utilities, which would enable cross-platform energy comparisons at all hardware tiers.

Low Complexity observations:

Total Tokens Per Total Joules:

At this low complexity, in the total tokens per total joules (Low Complexity - 1), the RTX 3070 came out on top, consuming less joules per tokens generated overall. At the 1.5 Million tokens generated mark, the RTX 4060 had spent 14794 joules while the RTX 3070 only spent 13227 joules.

Winner → RTX 3070: exhibits an 11.8% better performance compared to the RTX 4060.

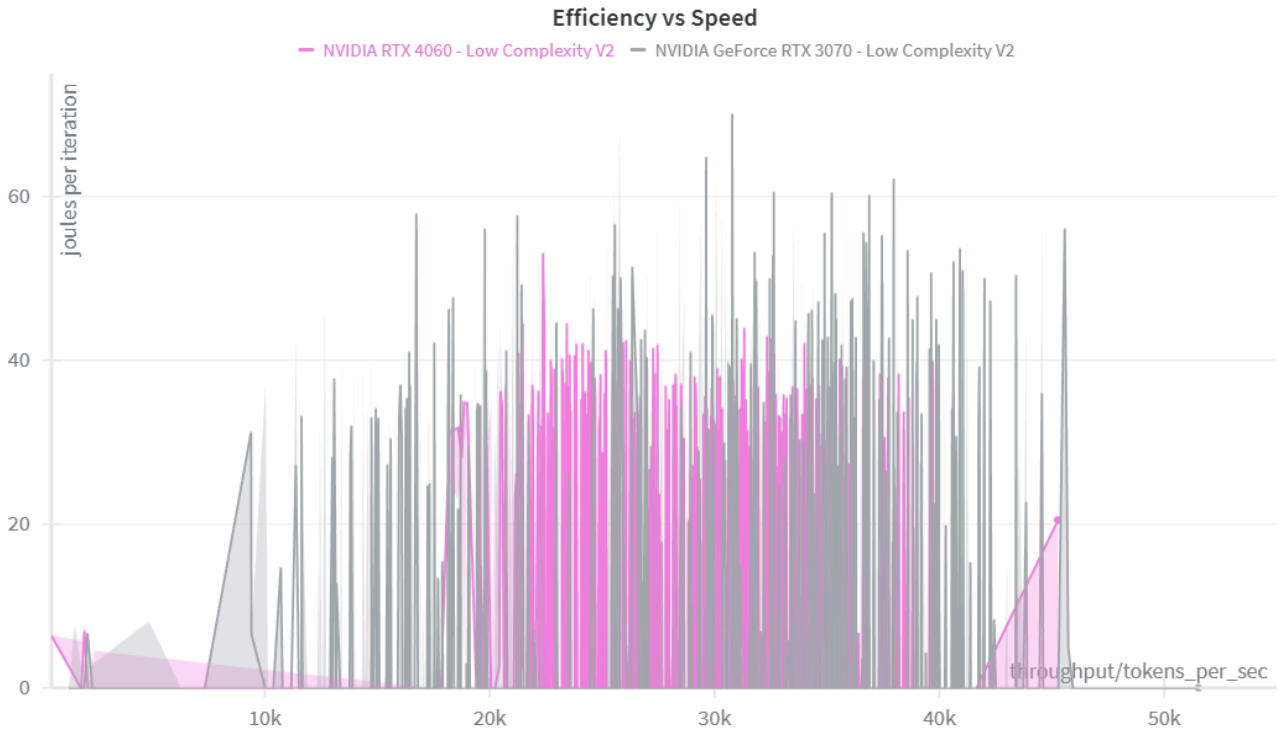


▲ Low Complexity - 1

Token Throughput vs. Joules per Iteration ("Efficiency vs. Speed"):

Both GPUs operate across a **comparable** throughput range of roughly 10k–50k tokens per second seen in graph Low Complexity - 2, so speed is not a differentiating factor here. The key distinction is energy consistency: the RTX 3070 produces frequent spikes reaching up to ~65 joules per iteration scattered across the full throughput range, while the RTX 4060 stays largely compressed between 20–40 joules with far fewer outliers. Since both cards move at similar speeds but the 4060 does so at a lower and more stable energy cost, the efficiency advantage is clear.

Winner → **RTX 4060**: equal throughput, meaningfully better energy consistency.



▲ Low Complexity - 2

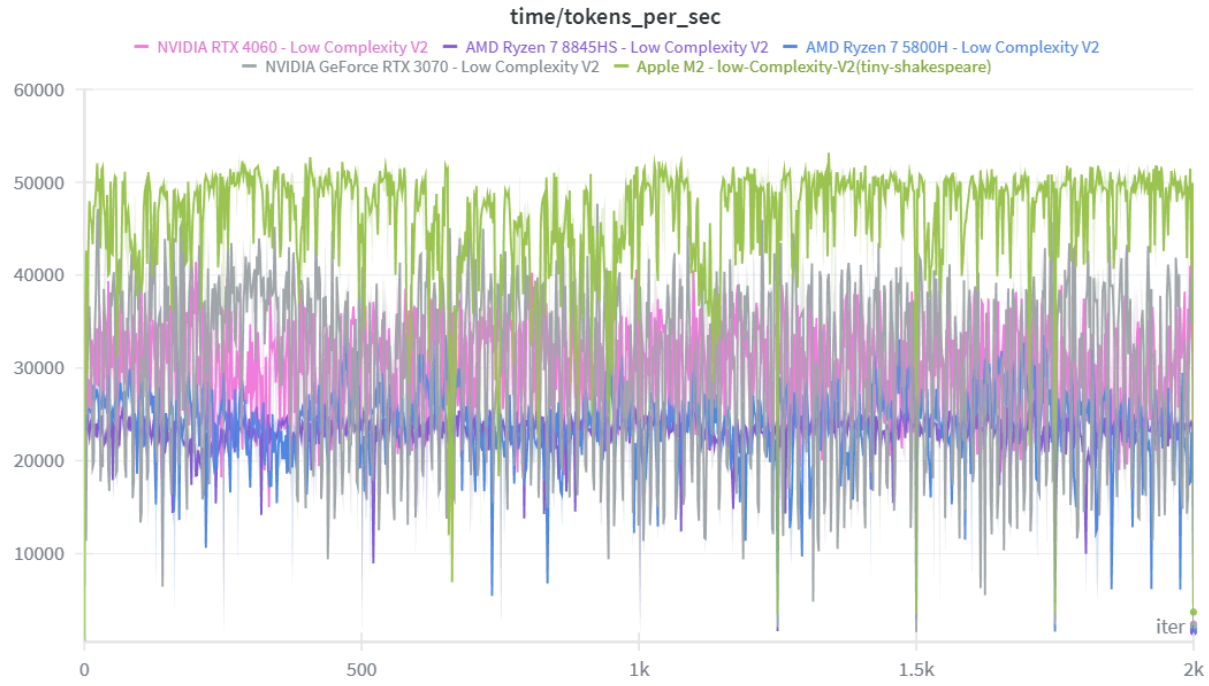
Tokens Per Second:

In the Low Complexity tokens per second graph, the Apple M2 demonstrates a clear advantage, averaging 45,000 to 50,000 tokens per second. This superior performance in lightweight tasks is largely attributed to its unified memory architecture, which allows for highly efficient data transfer with minimal latency.

Estimated Averages:

- Apple M2: ~40,000 - 50,000 tokens/sec
- NVIDIA RTX 4060: ~25,000 - 35,000 tokens/sec
- NVIDIA GeForce RTX 3070: ~25,000 - 30,000 tokens/sec
- AMD Ryzen CPUs (8845HS & 5800H): ~20,000 - 25,000 tokens/sec

Winner → **Apple M2**: highest overall throughput, dominating lightweight AI workloads due to its highly efficient unified memory architecture.

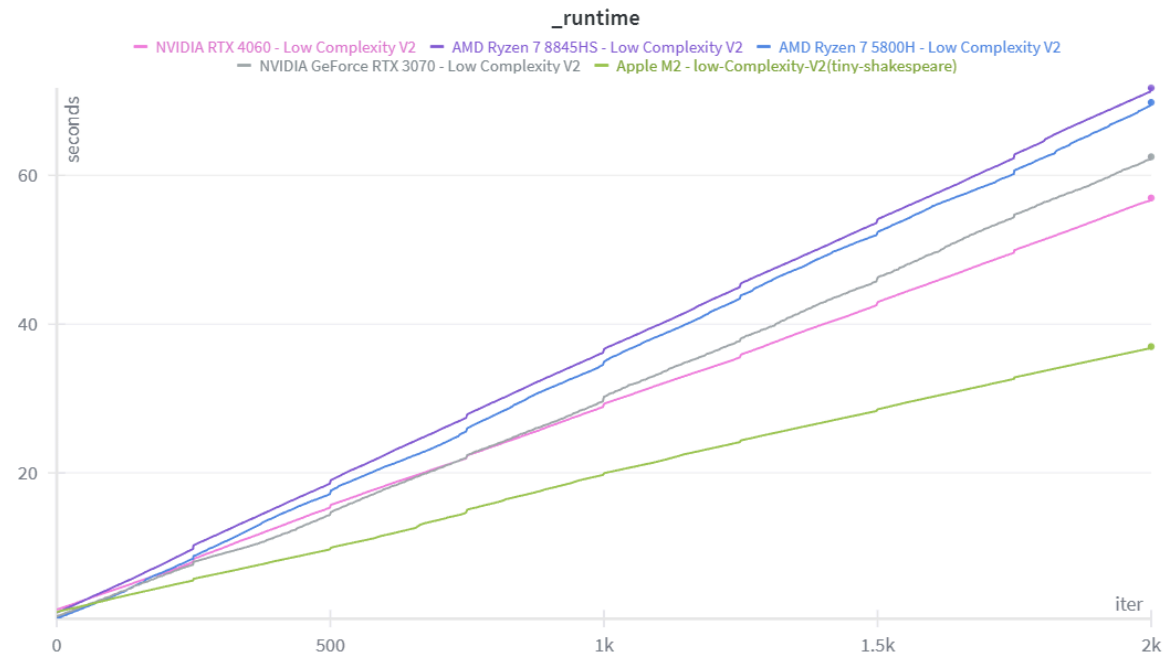


▲ Low Complexity - 3

Runtime:

For the low complexity, we can see that the fastest hardware was the Apple M2 chip then the RTX 4060, RTX 3070, AMD Ryzen 7 5800H, and AMD Ryzen 7 8845HS followed suit in that exact order. We can see that the M2 outperformed the GPUs probably due to this technicality that GPUs work better with larger workloads rather than small ones which are present at this low level of complexity.

Winner → Apple M2 Chip: Fastest hardware configuration



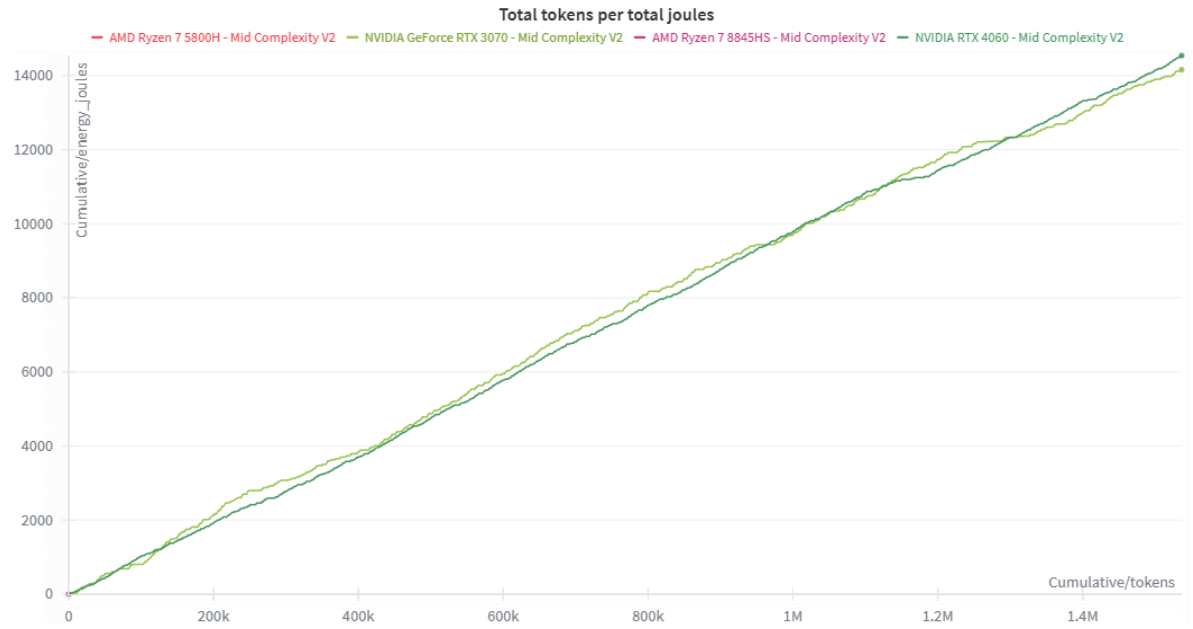
▲ Low Complexity - 4

Mid Complexity observations:

Total Tokens Per Total Joules:

The mid complexity graph of total tokens per total joules (Mid Complexity - 1) differed greatly from the low complexity one we just observed. In this instance, throughout all of the marks of tokens generated we can see that the RTX 4060 outperformed the RTX 3070 by a small amount. However, at some token generation marks the RTX 3070 was the one outperforming the RTX 4060. Therefore, we cannot truly say that one GPU outperformed the other, as the difference was trivial in this complexity level.

Winner → RTX 4060 & RTX 3070

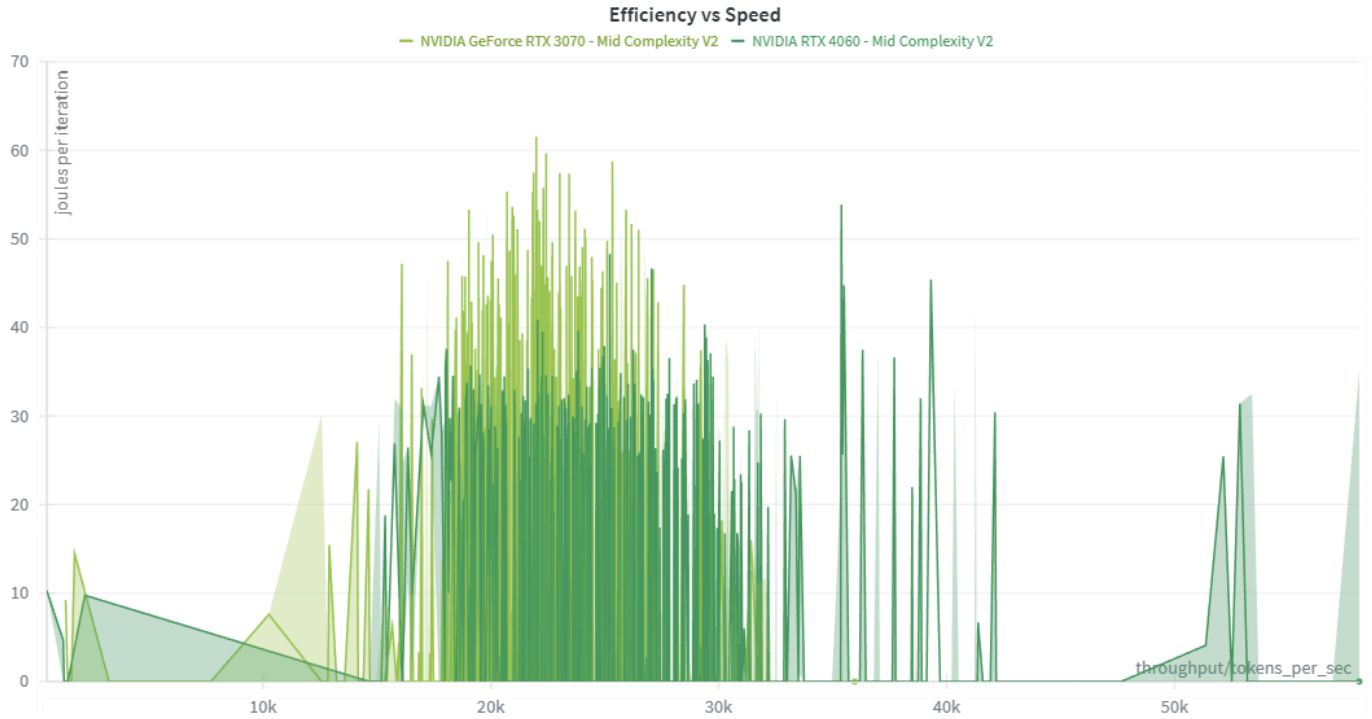


▲ Mid Complexity - 1

Token Throughput vs. Joules per Iteration ("Efficiency vs. Speed"):

This is the most competitive tier. Both GPUs reach similar peak throughput (~55k tokens/sec) and nearly identical peak joule values (~60–62 J) as seen in graph Mid Complexity - 2. The RTX 4060's high-density cluster sits slightly rightward on the x-axis, meaning it sustains its most energy-intensive iterations at marginally higher throughput by doing slightly more work for the same energy cost. The gap is narrow, but the pattern is consistent.

Winner → RTX 4060 (marginal): modest throughput edge at equivalent energy expenditure.



▲ Mid Complexity - 2

Tokens Per Second:

In the Mid Complexity tokens per second graph, a massive reshuffling of performance occurs as computational demands increase. The discrete GPUs take the lead, exposing the architectural limitations of standard CPUs and unified memory under heavier loads. The NVIDIA RTX 4060 maintains a strong baseline with significant performance spikes, while the RTX 3070 remains highly stable, highlighting the necessity of thousands of CUDA cores for more complex AI models.

Estimated Averages:

- NVIDIA RTX 4060: ~25,000 - 30,000 tokens/sec (with spikes up to 55,000+)
- NVIDIA GeForce RTX 3070: ~20,000 - 22,000 tokens/sec
- Apple M2: ~6,000 tokens/sec
- AMD Ryzen CPUs (8845HS & 5800H): ~2,500 - 3,500 tokens/sec

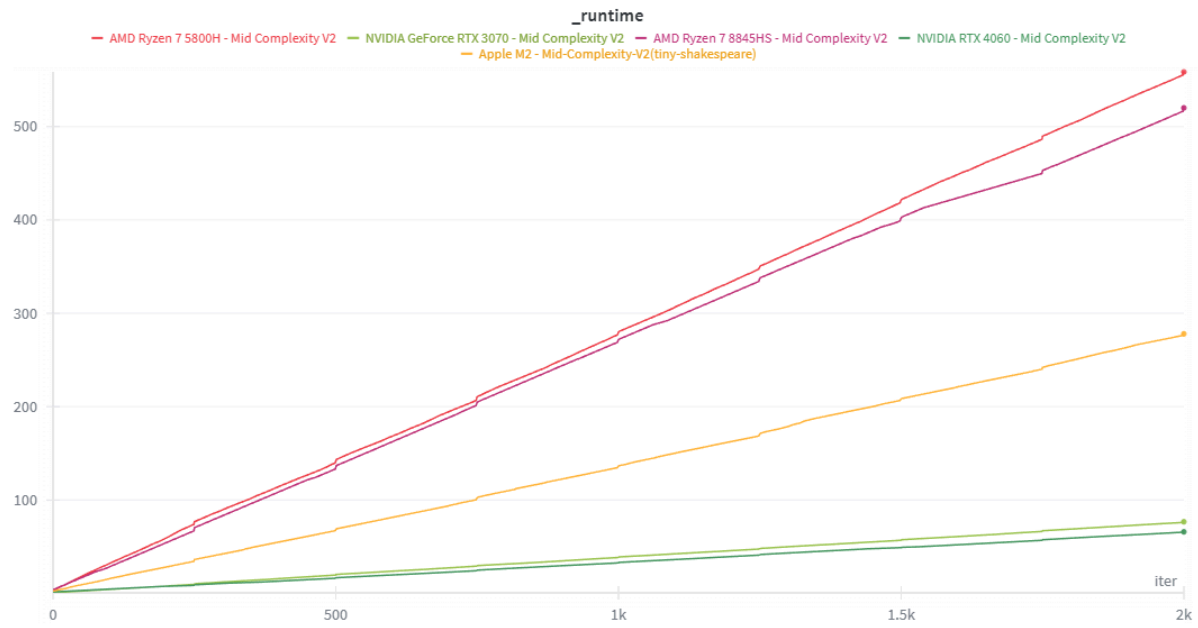
Winner: NVIDIA RTX 4060: highest peak throughput, successfully handling increased computational complexity better than CPU or unified memory architectures.

▲ Mid Complexity - 3

Runtime:

For the mid complexity, we can see that the fastest hardware was the RTX 4060 then the RTX 3070, Apple M2 Chip, AMD Ryzen 7 8845HS, and AMD Ryzen 7 5800H followed suit in that exact order. We can see that the M2 outperformed the AMD Ryzen CPUs because of its powerful more modern architecture. Also, the Apple M2 Chip has a unified memory system which allows it to share hardware resources more efficiently, increasing operational speed.

Winner → RTX 4060 (marginal): outperformed the RTX 3070 by 9 seconds.



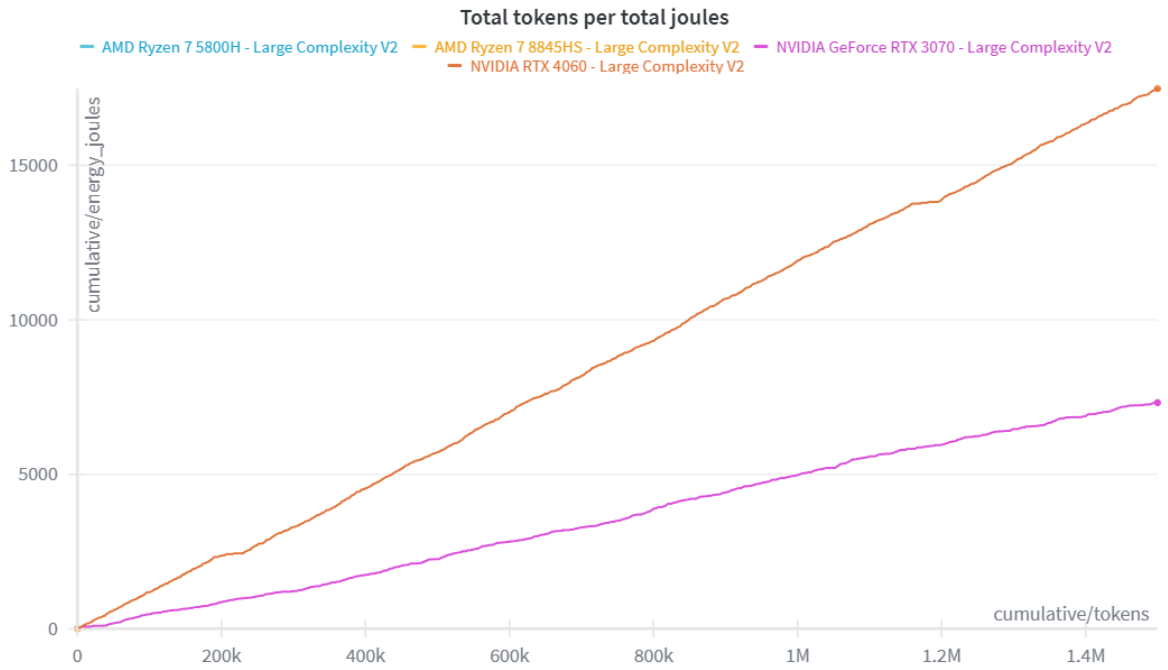
▲ Mid Complexity - 4

Large Complexity observations:

Total Tokens Per Total Joules:

For the large complexity we have a staggering difference in performance within the total tokens per total joules graph (Large Complexity - 1). From start to finish, the RTX 3070 outperformed the RTX 4060. When reaching the 1.5 Million tokens generated mark, the RTX 3070 had only spent 7315 joules compared to the RTX 4060's 17485 joules. This vast difference in performance can only be explained by the RTX 3070's higher number of CUDA cores.

Winner → RTX 3070: the RTX 3070 performed 239% more efficiently than the RTX 4060



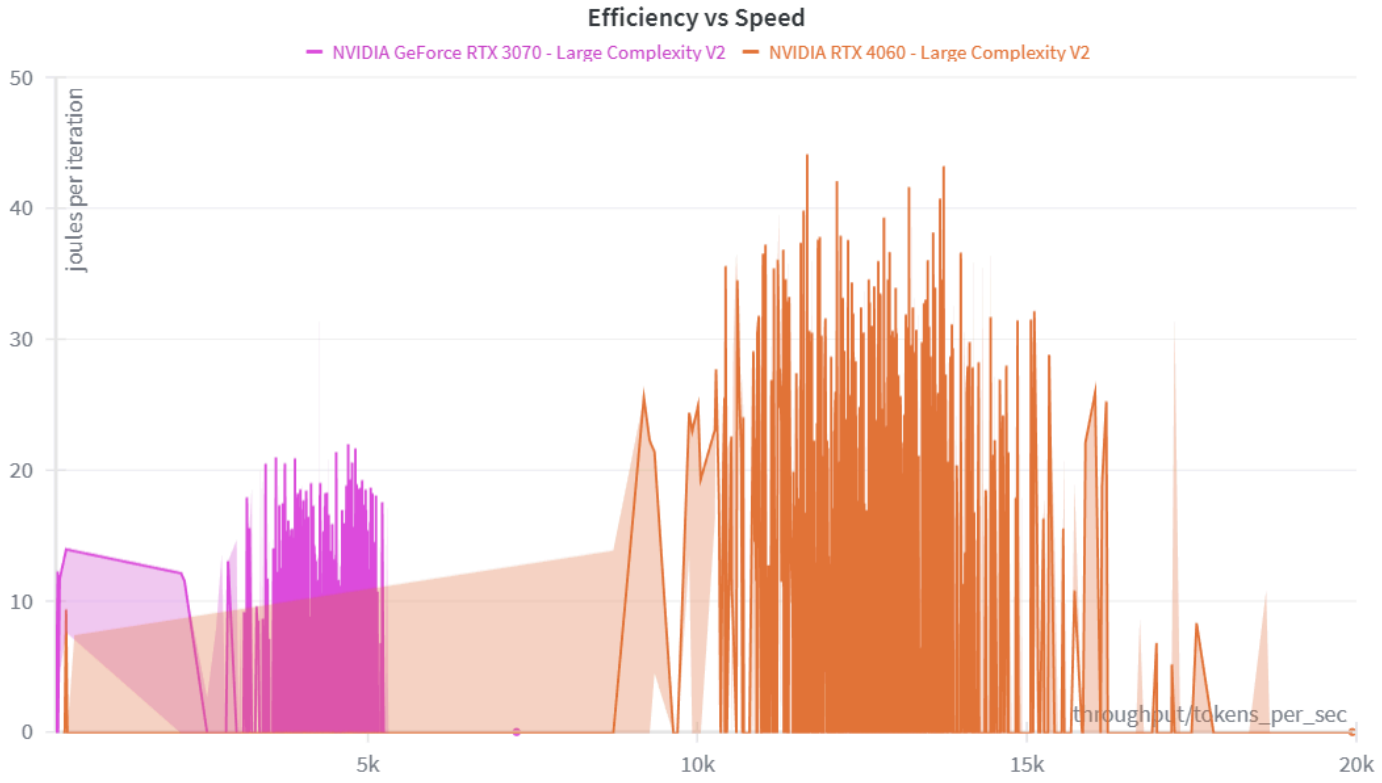
▲ Large Complexity - 1

Token Throughput vs. Joules per Iteration ("Efficiency vs. Speed"):

This is where the relationship inverts. The RTX 3070 is confined to roughly 1k–5k tokens per second, while the RTX 4060 operates at 8k–18k which is approximately 3 or 4 times faster as seen in graph Large Complexity - 2. However, the 4060 pays for that speed, with typical joule-per-iteration values of 25–42 J compared to the 3070's 10–20 J. The 3070 is slower but leaner; the 4060 is faster but energy-hungry.

- **For training speed:** RTX 4060 wins decisively.
- **For energy efficiency:** RTX 3070 wins at this scale.

Winner: Split as the 4060 dominates on throughput, the 3070 on joules per iteration.



▲ Large Complexity - 2

Tokens Per Second:

In the Large Complexity tokens per second graph, the workload is so demanding that overall token generation slows dramatically, necessitating a significantly lower y-axis scale (capped at 20,000). Under these high-intensity constraints, the NVIDIA RTX 4060 remains the undisputed leader. This demonstrates that for large-scale AI workloads, the extensive VRAM bandwidth and parallel computing architecture of modern discrete GPUs are strictly required to prevent severe system bottlenecks.

Estimated Averages:

- NVIDIA RTX 4060: ~12,000 - 13,000 tokens/sec
- NVIDIA GeForce RTX 3070: ~4,000 - 5,000 tokens/sec
- Apple M2: ~1,000 tokens/sec
- AMD Ryzen CPUs (8845HS & 5800H): < 500 tokens/sec

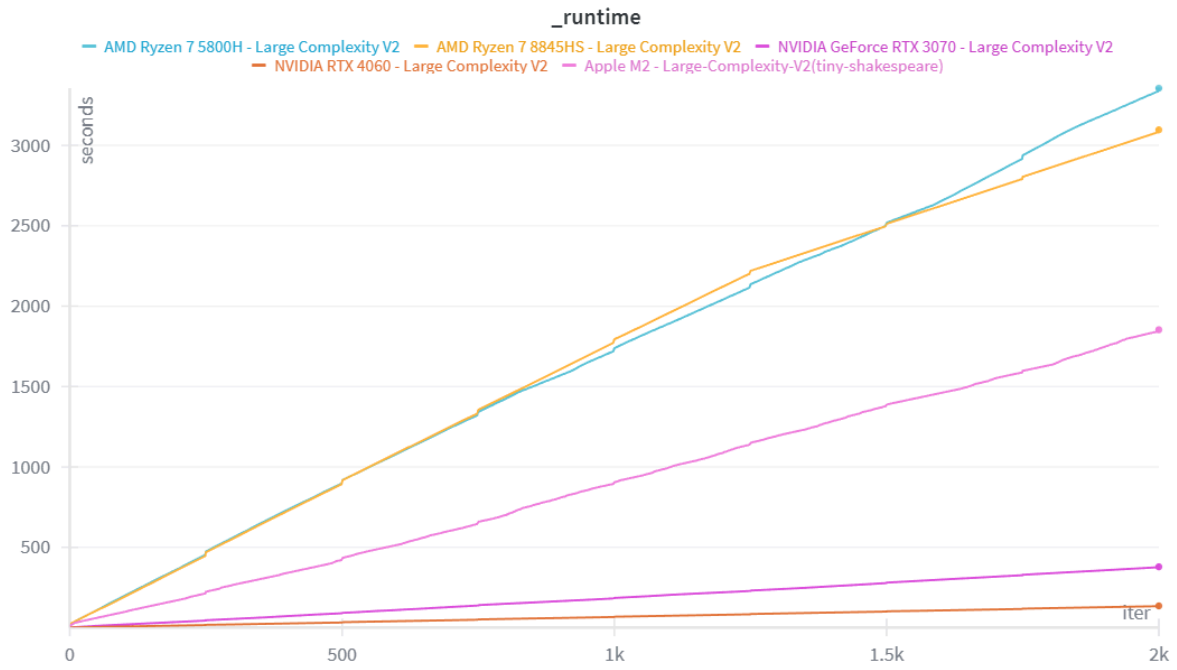
Winner → NVIDIA RTX 4060: undisputed throughput leader, maintaining viable performance under heavy computational loads where all other tested architectures severely bottleneck.

▲ Large Complexity - 3

Runtime:

For the large complexity, we can see that the fastest hardware was the RTX 4060 then the RTX 3070, Apple M2 Chip, AMD Ryzen 7 8845HS, and AMD Ryzen 7 5800H followed suit in that exact order. This time, the RTX 4060 had a significantly more efficient runtime than the RTX 3070 compared to the mid complexity level runtime figure we observed. We can see that the M2 outperformed the AMD Ryzen CPUs because of its powerful more modern architecture. Also, the Apple M2 Chip has a unified memory system which allows it to share hardware resources more efficiently, increasing operational speed.

Winner → RTX 4060: outperformed the RTX 3070 by 243 seconds.



▲ Large Complexity - 4

Next Steps:

- **Increase batch size**: the current fixed batch size of 12 is conservative. Increasing it will stress-test hardware memory bandwidth more aggressively and may reveal additional differentiation between the RTX 4060 and RTX 3070, particularly at large complexity where VRAM constraints become a limiting factor.
- **Try a larger dataset**: Tiny Shakespeare is a small, low-diversity corpus. Introducing a larger dataset such as OpenWebText would increase the number of training iterations required and provide a more realistic benchmark for sustained hardware performance over longer runs.

Conclusion:

This experiment set out to determine whether hardware specification differences produce observable variation in training behavior for a character-level GPT model trained on the Tiny Shakespeare dataset across three model complexity levels. The results confirm **two parallel findings**: training loss and validation loss remain effectively identical across all hardware configurations, validating experimental integrity, while hardware-sensitive metrics such as runtime, tokens per second, joules per iteration, and cumulative energy consumption do reveal substantial and consistent differentiation.

At **low complexity**, the Apple M2's unified memory architecture produces the highest token throughput, while the RTX 4060 demonstrates better energy consistency than the RTX 3070 on a per-iteration basis. As complexity scales to **mid level**, discrete GPUs assert dominance in throughput, with the RTX 4060 taking a marginal lead in both speed and efficiency. At **large complexity**, the most consequential inversion occurs: the RTX 4060 achieves 3 to 4 times higher throughput than the RTX 3070, but does so at roughly **239%** greater cumulative energy cost, while the Apple M2 and AMD Ryzen CPUs become effectively non-competitive under the demands of a 12-head, 12-layer, 768-dimensional model.

Taken together, these findings suggest that hardware selection for small-scale language model training involves a non-trivial tradeoff between speed and energy efficiency that is highly dependent on model scale. The RTX 4060's Ada Lovelace architecture is the most versatile performer across all three complexity levels, but the RTX 3070's superior energy efficiency at large scale demonstrates that newer hardware does not unconditionally outperform older generations on all relevant metrics.